Technical Appendix D

Locational Data for TRI Reporting Facilities and Off-site Facilities

Table of Contents

1. Introduction

The RSEI model uses latitude and longitude coordinates for each TRI reporting and off-site facility to fix each facility on the grid that underpins the model. The facility's location determines many of the modeling inputs, including the exposed population. With changes made to the model for Version 2.1, including more detailed air modeling close to the facility, and full-model results for surface water media, accurate locational data takes on additional importance.

There are two types of facilities included in the model, TRI reporting facilities and off-site facilities. The quality of locational data varies significantly between the two types. TRI reporters submit their own addresses and estimates of their latitude and longitude (lat/long) on Form R every year when they submit their release reports. These reports are subject to common reporting errors: transposition of digits, confusion of latitude with longitude, lack of precision, and nonreporting. The quality of reported data for off-site facilities is much worse, as the name and address of these off-site facilities are reported by the TRI reporters transferring the waste, not the receiving facility itself. The name and address tend to be reported in slightly different ways by different reporters, and often misspelled or misreported. Latitude and longitude are not reported at all. Little standardization is performed by TRI program, therefore minor differences in an off-site facility record, such as a slight misspelling of the name, or "St." instead of "Street", can make two records look like two different facilities, when they are really the same.

In RSEI Version 1.x, reporting facilities were located on the grid using their reported latitudes and longitudes, and off-site facilities were located using the coordinates of the centroid of their 5-digit ZIP code. For Version 2.1, the lat/longs for both reporters and off-site facilities were improved using a commercial geocoding service. Geocoding is a process where a computer program uses street address, city, state, and ZIP code to match addresses to geographic points in Census TIGER files, and then determines the latitude and longitude of the address. Each reporting facility also has submitted lat/long coordinates. In previous years, EPA has done some quality assurance (QA) work on those submitted coordinates; the process performed was one of determining the quality of each set of coordinates, and picking the highest quality set. For off-site facilities, the chief hurdle was to identify all of the different ways each true unique off-site facility could be reported, and then use the most accurate geocoded results for that unique facility.

Section 2 describes the geocoding process. Section 3 presents the geocoding results for TRI reporting facilities, and describes how the highest-quality final coordinates were determined. Section 4 presents the geocoding results for off-site facilities, and describes how the set of all reported off-site facilities were collapsed into a set of unique facilities with best coordinates.

2. Results of Geocoding

A commercial firm, Thomas Computing Services (TCS), was selected to perform the geocoding process. TCS geocodes data using the GDT software package Matchmaker 2000 version 2.3. This geocoding process involved matching records in the address databases to a reference street map. The reference street map with positioning information is based on the U.S. government TIGER census files.[1] Matchmaker links records in the two databases by matching street names and addresses. When the database records are successfully matched to a reference street map database, the record is considered a match and tagged with the correct latitude and longitude coordinates from the reference street map. After geocoding, some nonmatched records are matched manually, using Internet resources, other databases, and direct contact with the facility.

The matches are broken down into the following different types:

1. *Street segment exact match*- address is matched to a specific segment of a street, including matches that were made manually;
2. *ZIP+4 centroid match*- address is matched to a specific ZIP code plus 4 centroid;
3. *ZIP+2 centroid match*- address is matched to a specific ZIP code plus 2 centroid;
4. *ZIP code centroid match*- refers to the number of points matched only to a five-digit ZIP code centroid; and
5. *No match*- none of the above matches is detected.

TCS separately geocoded the TRI reporters and the off-site facilities. The results of each process are discussed in Sections 3 and 4, respectively.

[1] The data is based on public record and cannot be copyrighted, therefore it does not have licensing restraints.

3. TRI Reporting Facilities

The database of all TRI reporting facilities for 1988-2000 includes 45,651 facilities. As described above, these facilities were geocoded by TCS. The results are shown below in Table D-1.

Table D-1
Match Results for TRI Reporting Facilities

Coordinate Matches	Number of Records	Match Percentage
Street address (includes manual matches)	31,944	69.97%
ZIP+4 centroids	161	0.35%
ZIP+2 centroids	468	1.03%
ZIP code centroids	12,711	27.84%
Total matches	***45,284***	***99.20%***
Unmatched	***367***	***0.80%***

These results were combined with other data to determine the most accurate set of coordinates for each facility. The process and results are described below.

3.1 Inputs

Four tables were used in this analysis to determine the best set of coordinates:

1. **'Facility'**. This table contains the data on TRI reporting facilities from the May 2000 Public Data Release TRI data freeze, including the 45,651 facilities currently or historically reporting to TRI, up to and including TRI Reporting Year 2000. Fields relevant for this analysis are TRI ID (field name 'FacilityID'), county, state, submitted latitude and longitude (as submitted by the reporting facility), and preferred latitude and longitude (the facility-submitted coordinates after some annual quality assurance performed by EPA).

2. **'Pref94D'**. This table was provided by Loren Hall of U.S. EPA in dbf format. It contains the results of a QA process done in 1998 on TRI reporting data up to and including RY1996. Data includes, for 36,652 facilities, submitted lat/longs, preferred lat/longs, codes describing the

level of accuracy of the preferred lat/longs (in field PREFER_AC), and the QA checks themselves that were done in order to determine the quality of the lat/longs that were considered "preferred" (PREFER_QA). There is a preferred lat/long for each facility, and some of them failed very basic tests, so "preferred" should simply be taken as a sign that a set of coordinates went through tests, not that they are necessarily of a high quality. The quality of the preferred lat/longs can only be determined by looking at the fields PREFER_QA and PREFER_AC. See below for details.

3. **'Old_gdt'**. This table is based on a Lotus 1-2-3 worksheet, 'trigdt', also provided by Loren Hall. This file was generated by OIRM's (Office of Information Resources Management) System Development Center in 1998, and was a geocoding effort of all regulated entities known to EPA (some 25 million addresses). In this exercise, duplicate facility records were not eliminated, since the Agency did not want to miss any changed addresses. This geocoded and Q/C'd data became the source of most of the data in the Envirofacts database. The file contained 17,286 records (after records with zero lat/longs were deleted), with geocoded results (lat/long, confidence level information) for each one. In this Appendix, this dataset will be referred to as the old gdt dataset.

4. **'New_gdt'**. This table is based on data provided by TCS, the private company contracted to geocode reporting and off-site facilities. The file contains all of the current and historically reporting facilities that were geocoded to differing levels of precision, from street segment address matches to 5-digit ZIP code matches. Because the data set that TCS geocoded was simply a later version (with some extra new reporters) of the data that was geocoded in 1998, the results were compared and found to be very similar. This is because TCS uses the same gdt software used by OIRM. In this Appendix, this dataset will be referred to as the new gdt dataset.

3.2 Overview of Process

In consultation with Loren Hall, a basic process to update low-confidence lat/longs with geocoded data was determined:

1. Combine the old and new gdt geocoded data to create one dataset;

2. Extract the high-quality preferred lat/longs from pref94D, and preserve them for the final dataset;

3. Compare the lat/longs from the remaining TRI facilities with the lat/longs from the geocoded dataset, calculating the distance between the two sets of coordinates; and

4. In a series of steps, replace the TRI lat/longs with the geocoded lat/longs if a) the distance between the two is greater than a determined minimum, and b) the confidence level of the geocoded lat/longs is above a certain minimum.

3.3 Details of Process

This analysis was conducted in Access. Each step of the process is described in the sections below.

3.3.1 Combine the TCS and the gdt geocoded data to create one dataset

There is a large degree of overlap between the new and old gdt datasets, with approximately 17,000 facilities in common. The coordinates of the duplicates facilities were compared. If there was no difference, then the new gdt fields were adopted, as the more current source. If there was a difference, then the coordinates were compared based on the strength of the geocoding. The values in the following fields were compared:

Table D-2
Match Level Codes
(In descending order of quality)

Field 'XIN'	Field 'STAT'	Type of Match
0, S, V or I	B1, R1, B2, R2, B3, R3, R4, B5, R5,	matched to a street segment
4		matched to a 4-digit ZIP code centroid
0	B6, R6, B7, R7	matched to a placeholder
2		matched to a 2-digit ZIP code centroid
X		matched to a 5-digit ZIP code centroid;

An XIN = 0, STAT = B1, R1 match is the most accurate, and a XIN = 5 is the least accurate. The most accurate match was chosen for each facility. In the cases where the confidence level was the same, the new gdt coordinates were chosen as the most recent source.

The duplicate facilities described above were combined with the unique facilities from each dataset into one table containing one record for each geocoded facility with the best set of geocoded coordinates.

3.3.2 Extract the high-quality preferred lat/longs from pref94D, and preserve them for the final dataset

In order to preserve the high-quality lat/longs from the round of quality checks performed in 1998, those considered 'high quality' were extracted from the table 'pref94D' and set aside. The following criteria for 'high quality' were developed in consultation with Loren Hall:

1. Facilities where the third position of PREFER_QA field is "1"(indicating that the submitted coordinates were found to be within 2 km of a reasonably good alternate coordinate value); or

2. Facilities where the fourth position of PREFER_QA field is "V", "A", or "D" (manually verified to have preferred coordinates); or

3. Facilities where the PREFER_AC field value is <150 meters (the coordinates are considered accurate to within 150 meters, based on the kind of check performed).

A set of preferred lat/longs meeting any one of these conditions was considered high quality and set aside. There were 18,036 facilities originally in this group. Sixty-three records with zero values in the PREFER_AC (which describes in meters the level of accuracy of the preferred lat/longs) were deleted from this set, and therefore went through the rest of the process like other non-high quality lat/longs. Deleting them left 17,973 facilities designated as 'High Quality.' The remaining 27,678 facilities without high-quality TRI coordinates then went through the comparison with the geocoded facilities described in the next section.

3.3.3 Compare the lat/longs from the remaining facilities in pref94D with the lat/longs from the geocoded dataset, calculating the distance between the two sets of coordinates

The facilities without high-quality TRI coordinates were matched against the facilities in the combined gdt dataset. For each facility in this set, it was necessary to select a lat/long to compare with the geocoded lat/long. There are two possibilities in each dataset (the Pref94D database or the current set of TRI reporters called 'Facility'): the Preferred lat/long, or the Submitted lat/long. Following further consultation with Loren Hall, the decision was made to compare the geocoded results with the Submitted rather than the Preferred coordinates. This comparison was considered more appropriate because a number of instances were identified where one of the QA tests used to derive Preferred values erroneously rejected valid submitted coordinates. Therefore, using the Submitted coordinates will avoid perpetuating a situation where EPA rejected a valid submitted coordinate. Loren Hall also advised that it was better to use the submitted coordinates from the 1998 data set (Pref94D), because in some instances erroneous Preferred coordinates had been entered into facilities' Form R's before they were sent out to them (to simplify reporting for the facilities). In many cases where these facilities Preferred coordinates were wrong, the error was probably not discovered. This may have resulted in the erroneous Preferred coordinates becoming erroneous Submitted coordinates, thereby perpetuating the error.

Using the logic described above, each facility was assigned its 'best' set of TRI coordinates (TRI lat/long), using the following hierarchy:

- 1994 Submitted (from 'Pref94D');
- 2000 Submitted (from 'Facility');
- 1994 Preferred (from 'Pref94D');
- 2000 Preferred (from 'Facility').

The distance (in kilometers) from each facility's best TRI coordinate to its best gdt (geocoded) coordinate was then calculated using the following formula:

Distance = 6377*acos(cos(rad(90-TRI Lat))*cos(rad(90-gdt Lat)) + sin(rad(90-TRI Lat)) *
sin(rad(90-gdt Lat)) * cos(rad(TRI Long*-1) - gdt Long)))

The resulting value was then used in determining whether to retain the TRI coordinates or substitute the gdt coordinates, as described below.

3.3.4 In a series of steps, replace the low-quality preferred lat/longs with the geocoded lat/longs

The basic premise of the following steps is that one can have greater confidence replacing low-quality lat/longs with geocoded lat/longs if the confidence level associated with the geocoded lat/longs is very high, and the distance between the two sets of coordinates is very great. In these cases one can feel confident that the submitted lat/long is simply erroneous. As the confidence level of the geocoded lat/longs decreases toward the level of a ZIP code centroid, one cannot be sure that differences of a few kilometers do not simply represent real distances from a plant to the centroid of its ZIP code. Therefore, the replacement of the low-quality preferred lat/longs was done in a series of steps that accounted for both the distance between the sets of coordinates and the confidence level of the geocoded results.

At each step, the geocoded coordinates that matched the criteria were substituted for the TRI coordinates and set aside. For the field 'Final Source,' the following codes were used:

1. QA_GDT. This code refers to when geocoded lat/longs from the combined gdt database were substituted for the TRI lat/longs.
2. QA_TRI. This code refers to when the pairs of lat/longs did not meet any of the criteria above, so the TRI Form R-reported lat/longs were retained.

Step 1. Low-quality lat/longs were replaced with geocoded coordinates if the distance between the two sets of coordinates was greater than or equal to 2 km, AND the geocoded result was matched at a street segment or intersection– i.e., the GSTAT field showed B1, B2, B3, B5, R1, R2, R3, or R4 and the GDTXIN field showed 0 or S, V or I.. In this step, 8304 facilities were assigned the code 'QA_GDT,' and 8208 facilities were assigned the code 'QA_TRI'.

Step 2. Low-quality lat/longs were NOT replaced with geocoded coordinates if the distance between the two sets of coordinates was greater than or equal to 2 km and less than 5 km, AND the geocoded result was matched at a street segment or intersection– i.e., the GSTAT field showed B1, B2, B3, B5, R1, R2, R3, or R4, AND the facility reported MORE THAN 1,000,000 pounds of total releases, including direct releases and off-site transfers. TRI release data from 1999 was used for this test, except for facilities new to TRI in 2000, for which 2000 data was used. In these cases we are assuming that the large plants know their locations well and may have a good reason to report addresses up to 5 km different from their lat/longs; for instance, the geocoded result may represent the 'front door' of the facility, but the submitted lat/long represents either the point of release or the center of production. In Step 2, we are assuming that facilities releasing less than 1,000,000 pounds in 1998 are not large enough to have such an issue, and that the difference in the coordinates represents a lack of precision on their part. In this step, 48 facilities were assigned 'QA_ TRI'.

Step 3. Low-quality lat/longs were replaced with geocoded coordinates if the distance between the two sets of coordinates was greater than or equal to 10 km, AND the geocoded result was matched at the ZIP+4 centroid level. In this step, 47 facilities were assigned to 'QA_GDT' and 208 were assigned 'QA_TRI'.

Step 4. Low-quality lat/longs were replaced with geocoded coordinates if 1) the distance between the two sets of coordinates was greater than or equal to 15 km, AND 2) the geocoded result was matched at either the ZIP+2 centroid level OR the B6/R6 placeholder match. In this step, 142 facilities were assigned to 'QA_GDT' and 837 were assigned 'QA_TRI'.

Step 5. Low-quality lat/longs were replaced with geocoded coordinates if the distance between the two sets of coordinates was greater than or equal to 20 km, AND the geocoded result was matched at the 5-digit ZIP code centroid level. In this step, 1386 facilities were assigned to 'QA_GDT' and 8121 were assigned 'QA_TRI'.

3.3.5 Replace the reported county name with the geocoded name for the county in any instances where the field is now blank

After the final table was created (see below), county names and FIPS codes were pulled in from the 'Facility' table. The county field was checked for blanks, but no blanks were found in RY 2000.

3.3.6 Facilities with missing coordinates and quality assurance

As described earlier, there were 108 facilities which had geocoded results, but did not have valid (nonzero) TRI coordinates for comparison. These facilities were assigned the source code, 'GDT_NOTRI,' and the coordinates were taken from the new gdt table.

There were 403 facilities in the set of facilities without high-quality TRI coordinates that were not found in the combined set of gdt facilities, for the most part because the facilities were located in places other than the fifty U.S. States and the District of Columbia. These facilities were assigned the source code, 'TRI_NOGDT,' and the coordinates were taken from TRI in the hierarchy described earlier. That left 32 facilities still without coordinates. Of these 32 facilities, 18 had coordinates in the high-quality preferred set, so these were simply deleted. That left 14 facilities with neither TRI nor GDT coordinates. All 14 of these facilities have TRI values for 1999, so these values were used. However, these values in 1999 were taken from the coordinates submitted in 1998, so the code '1998TRI' was assigned.

Once all of the facilities had been assigned coordinates and pulled together in a draft table, the final coordinates were plotted in a GIS (Geographic Information System) program. The state that the coordinates were plotted to were matched against the reported state and visually inspected for those that did not match. Those that fell on a coast or a river state boundary were considered allowable, and those on straight state boundaries were given a one-mile tolerance before being counted as incorrect. The final result was 30 facilities that plotted outside of their reported state. Seventeen additional errors in the vicinity of PR were added. Using the TRI hierarchy and the gdt coordinates, if available, additional coordinates were checked for each of these failed facilities. If not valid coordinates could be found in the TRI or gdt data, EPA's LRT system was checked. If that also failed, the facility's reported zip code centroid was adopted, using an internet-based zip-code lookup.

Additionally, the coordinates for three facilities that had been previously checked using geocoding and maps were provided by Loren Hall of U.S. EPA. These facilities were also assigned the code 'MANUAL,' and changed by hand in the Draft final table. These three facilities are listed in Table D-3.

Table D-3.
Facilities with Coordinates Corrected After Mapping

Facility Id	Original Lat	Original Long	Original Source	Final_lat	Final_long	Final_source code	Data Source
46517LRMDW58288				41.645432	-85.991839	MANUAL	EPA (Provided by Loren Hall, map look-up.)
46517LRNC 28858				41.648419	-86.019967	MANUAL	EPA (Provided by Loren Hall, map look-up.)
46517TTFRM28816				41.648398	-86.019364	MANUAL	EPA (Provided by Loren Hall, map look-up.)

3.3.7 Creating the Final Table

The final coordinates from the comparison process described in section 3.3.4 were combined with the original High-Quality coordinates that were set aside in section 3.3.2.

Tables D-4 and D-5 show the data fields that will be added to the 'Facility' table used in the RSEI model, and how the contents of each field were derived. These tables do not include the Method, Accuracy, and Description (MAD) codes, which are described separately in Section 3.4, below.

Table D-4
New Data Fields in Final Table

Field	Derivation of Contents
Facility ID	Facility ID used in TRI reporting, unique for each facility.
SubLat SubLong	Coordinates originally submitted by each reporting facility as reported in the current year TRI data freeze. **NOTE:** This field may not match the submitted coordinates used to map the facility in the RSEI model, as the submitted coordinates are taken preferentially from the internal EPA dataset 'Pref94D.'
PreferLat PreferLong	QA'd coordinates derived by EPA for each reporting facility as listed in the current year TRI data freeze. **NOTE:** This field may not match the preferred coordinates used to map the facility in the RSEI model, as the submitted coordinates are taken preferentially from the internal EPA dataset 'Pref94D.'
Latitude Longitude	These are the final coordinates that will be used in the RSEI model. Their derivation depends on what is in the 'FINAL_SOURCE' field (see below).
LatLongSource	**QA_GDT**. Ultimate source is geocoded data using GDT software, performed either by EPA in 1998 or by TCS in the current year for EPA. Substituted for low-quality TRI lat/longs (see Section 3.4) **QA_TRI**. Ultimate source is TRI, as reported either in 'Pref94D' or in the 'Facility' table in the current year TRI data freeze. Coordinates could not be replaced by geocoded results (see Section 3.4). **HQPREFER**. Ultimate source is 'PREF_LAT' and 'PREF_LONG' in 'Pref94D'. Originally selected as High Quality Preferred Lat/longs; no comparison to geocoded results was performed. **TRI_NOGDT**. Ultimate source is TRI, as reported either in 'Pref94D' or in the 'Facility' table in the current year TRI data freeze. One of approximately 400 facilities with TRI coordinates that did not have geocoding results to compare against. **GDT_NOTRI**. Ultimate source is geocoded data using GDT software, either by EPA in 1998 or by TCS in the current year for EPA. Adopted without comparison because no TRI coordinates were available.

Table D-5
Summary of Final Facility Coordinates

Code Used	Type of Match/Comparison	Source	Description	Num. of Facilities
QA_GDT	Street Address Match	GDT	TRI coordinates replaced with Geocoded	8254
QA_GDT	Zip+4	GDT	TRI coordinates replaced with Geocoded	47
QA_GDT	Zip+2 (or B6/R6)	GDT	TRI coordinates replaced with Geocoded	140
QA_GDT	5-digit Zip	GDT	TRI coordinates replaced with Geocoded	1390
QA_GDT	map plotting	GDT	GDT adopted after plot failure	4
HQPREFER	Within 2 km of alternate, manually verified, or accuracy within 150 km	Pref94D	TRI Preferred coordinates retained	17,899
QA_TRI	Step 1 (<2 km from geocoded address match)	Submitted*	TRI Submitted coordinates retained, after check against geocoded coordinates	8208
QA_TRI	Step 2 (2-5 km from geocoded address match, >1,000,000 lbs)	Submitted*	TRI Submitted coordinates retained, after check against geocoded coordinates	48
QA_TRI	Step 3 (<10 km from zip+4 geocoded match)	Submitted*	TRI Submitted coordinates retained, after check against geocoded coordinates	206
QA_TRI	Step 4 (<15 km from zip+2 or B6/R6 geocoded match)	Submitted*	TRI Submitted coordinates retained, after check against geocoded coordinates	836
QA_TRI	Step 5 (<20 km from 5-digit zip code match)	Submitted*	TRI Submitted coordinates retained, after check against geocoded coordinates	8,116
QA_TRI	map plotting		TRI adopted after plot failure	17
TRI_NOGDT		Submitted*	TRI Submitted coordinates retained; no geocoded results available for comparison	365
GDT_NOTRI		GDT	Geocoded coordinates used; no TRI coordinates available for comparison.	94
ZIP	Coordinates revised after map plotting	from zip code lookup	Zip code centroid from Internet zip code look-up used.	9
MANUAL	Coordinates revised after map plotting		Coordinates either mapped or coordinates adjusted (e.g.,, decimal place moved).	3
LRT	Coordinates revised after map plotting		Best Value from EPA's Locational Reference Table (LRT) System used.	1
1998TRI		1998 Submitted	Used 1998 TRI Submitted coordinates (no GDT or 1999 TRI coordinates available).	14

* Submitted coordinates were preferentially taken from 'Pref94D'; if not available there, they were taken from the 'Facility' table.

3.4 Method, Accuracy, and Description (MAD) Codes

Method, Accuracy, and Description (MAD) codes are standardized codes that describe how a set of lat/longs were generated, what quality assurance checks were performed on it, and how accurate it is considered to be. The codes allow for comparison of different sets of coordinates that were generated at different times and by different processes. These codes are used by EPA offices, contractors, and by EPA's centralized Locational Reference Table (LRT).

The 'Facility' table in the RSEI database contains some information on MAD codes. In some cases, for instance when the coordinates from EPA's 1998 QA process were adopted, the MAD codes already assigned were simply carried over and adopted. In other cases, where coordinates were adopted as a result of the comparisons described above, new MAD codes were assigned and added to the table. However, due to resource and time constraints, in some cases not all of the codes could be filled in. Table D-6 describes each of the MAD codes included in the 'Facility' table. This information is taken from a table provided by Loren Hall of U.S. EPA.

Table D-6
Description of MAD Codes in ‘Facility’ Table

Code	Length	Type	Description	Values	
PREFER_AC	8.2	N	Accuracy of the preferred coordinates (in m)		
PREFER_CM	2	C	Collection method code for the preferred coordinate (as specified in MAD code)	A1	Address matching-house number
				A2	Address matching-block face
				C2	Census block/group-1990-centroid
				C3	Census block tract-1990-centroid
				G3	GPS code measurements (pseudo range) differential (DGPS)
				G4	GPS code measurements (pseudo range) precise positioning service
				I1	Interpolation-map
				I2	Interpolation-photo
				OT	Other
				Z1	ZIPcode-centroid
				UN	Unknown

Table D-6
Description of MAD Codes in ‘Facility’ Table

Code	Length	Type	Description	Values	
PREFER_DC	2	C	Description category of the preferred coordinate (as specified in MAD code)	PG	Plant entrance (general)
				FC	Facility centroid
				CE	Center of facility
				OT	Other (Describe or name in description comments)
				UN	Unknown
REFER_HD	1	C	Horizontal datum of the preferred coordinate (as specified in MAD code)	1	NAD27
				2	NAD83
				O	Other
				U	Unknown
PREFER_SMS	1	C	Source map scale of the preferred coordinate (as specified in MAD code)	E	1:24,000
				J	1:100,000

Table D-6
Description of MAD Codes in 'Facility' Table

Code	Length	Type	Description	Values
PREFER_QA	4	C	Results of four quality assurance tests. It follows the current format for PREFERRED-QA-CODE in TRIS-PREFERRED-LOCATION, except for the fourth position.	First position: Point location was checked against ZIP code polygon of ZIP in address field or TRI facility ID (approximated by a rectangle with an additional 2 km buffer surrounding it): 0 Test was not performed 1 Test was performed and coordinates passed 2 Test was performed and coordinates failed
				Second position: Point location was checked against 25 km radius of ZIP code centroid of ZIP in address field or TRI facility ID (generally performed only if ZIP polygon test was not possible or likely to yield erroneous results, e.g. for rural ZIP codes): 0 Test was not performed 1 Test was performed and coordinates passed 2 Test was performed and coordinates failed
				Third position: Point location was compared to an alternate coordinate of known accuracy (e.g. below about 600m). If alternate coordinates were located within a 2 km buffer of submitted coordinates, the latter were accepted and assigned the estimated accuracy of the alternate coordinates. If the alternate coordinates were outside the buffer, the alternates were selected. If an alternate coordinate was selected as preferred, it should always have a value of 1 (while the corresponding submitted coordinates would have a value of 2). 0 Test was not performed 1 Test was performed and coordinates passed 2 Test was performed and coordinates failed

Table D-6
Description of MAD Codes in 'Facility' Table

Code	Length	Type	Description	Values
PREFER_QA cont.				The fourth position contains one of the following five values: 0 No manual verification was done V Manual verification was done I Manual verification was done, and no preferred coordinate could be selected A Manual verification was done, and its result agrees with the preferred coordinate generated by the automated selection process D Verification was done, and its result disagrees with the preferred coordinate generated by the automated selection process (manually verified coordinate was selected as the preferred value)
PREFER_MV	23	C	Results of verification.	This is a new field for the 1987-94 data and can contain results of up to six latitude and longitude verifications by EPA staff, grantees, or contractors through the given process(es). Its length is 23 alphanumeric (six 3 character segments, colon delimited. Please refer to Appendix A for detailed values.

Table D-7
Summary of Final Source Codes and MAD Codes Assigned

Final_Source Code Used	Type of Match/Comparison	Source	Description	MAD CODES

QA_GDT	Street Address Match	GDT	TRI coordinates replaced with Geocoded	AC= 150
				CM=A2

QA_GDT	ZIP+4	GDT	TRI coordinates replaced with Geocoded	AC = 4000
				CM=Z1

QA_GDT	ZIP+2 (or B6/R6)	GDT	TRI coordinates replaced with Geocoded	AC = 8000*
				CM=Z1

QA_GDT	5-digit ZIP	GDT	TRI coordinates replaced with Geocoded	AC = 11000
				CM=Z1

QA_GDT	Street Match	TCS	TRI coordinates replaced with Geocoded	AC= 150
				CM=A2

QA_GDT	ZIP+4	TCS	TRI coordinates replaced with Geocoded	AC = 4000
				CM=Z1

QA_GDT	ZIP+2 (or B6/R6)	TCS	TRI coordinates replaced with Geocoded	AC = 8000*
				CM=Z1

QA_GDT	5-digit ZIP	TCS	TRI coordinates replaced with Geocoded	AC = 11000
				CM=Z1

HQ_PREFER	Within 2 km of alternate, manually verified, or accuracy within 150 km	Pref94D	TRI Preferred coordinates retained	Maintain existing MAD codes in Pref94D (all 7 codes)

Table D-7
Summary of Final Source Codes and MAD Codes Assigned

Final_Source Code Used	Type of Match/Comparison	Source	Description	MAD CODES

QA_TRI	Step 1 (<2 km from geocoded address match)	Submitted	TRI Submitted coordinates retained, after check against geocoded coordinates	MV = A4H
				CM=A2

QA_TRI	Step 2b (2-5 km from geocoded address match, >1,000,000 lbs)	Submitted	TRI Submitted coordinates retained, after check against geocoded coordinates	MV = A5M
				CM=A2

QA_TRI	Step 3 (<10 km from ZIP+4 geocoded match)	Submitted	TRI Submitted coordinates retained, after check against geocoded coordinates	MV = Z6H
				CM=Z1

QA_TRI	Step 4 (<15 km from ZIP+2 or B6/R6 geocoded match)	Submitted	TRI Submitted coordinates retained, after check against geocoded coordinates	MV = Z7M
				CM=Z1

QA_TRI	Step 5 (<20 km from 5-digit ZIP code match)	Submitted	TRI Submitted coordinates retained, after check against geocoded coordinates	MV = Z7L
				CM=Z1

GDT_NOTRI	None	Submitted	TRI Submitted coordinates retained; no geocoded results available for comparison	NO MAD CODES
GDT_NOTRI	None	GDT	Geocoded coordinates adopted; no TRI coordinates available for comparison	Same MAD Codes as QA_GDT

* No information available accuracy of ZIP+2. Value is the rounded average of 5-digit ZIP and ZIP+4.

4. Off-site Facilities

Previously, all off-site facilities had been located on the model grid using the centroid of the facility's ZIP code. The geocoding effort represents a significant improvement from that methodology. However, the problems with the set of off-site facilities are longstanding and serious: most notably that unique IDs are not used by TRI, and the addresses are not reported by the facilities themselves, but by those facilities that transfer waste to them. Given this, the accuracy of the reported addresses is questionable. In addition, because many different reporting facilities may be transferring their waste to the same facilities, there are many instances of the same facility being reported with many different permutations of name and address. The biggest challenge in this exercise was to collapse the entire set of off-site facilities into a set of unique facilities. Briefly, the entire set of off-site facilities was geocoded by TCS, and then the whole set was run through a series of matching programs in SAS, designed to match facilities to each other, on name first (based on the assumption that a third party is most likely to get a facility's name correct), providing leeway for non-exact matches, and then moving through the rest of the facility's address and determining if it is a plausible match. After this exercise, the set of unique facilities was pared down from approximately 3 million to 39,000. Approximately 36% of these facilities had an address match more accurate than a ZIP+4 match.

4.1 Overview of Process

There are several data processing steps in determining unique facilities and their coordinates. First, in order to best determine unique facilities, the facility records were collapsed from approximately 3 million to almost 300,000 by removing the exact duplicates. Second, in order to expedite and improve the off-site facility locating process, TCS geocoded the data and reported match rates. Finally, the geocoded off-site facility data was further collapsed in order to remove non-exact duplicates and determine truly unique off-site facilities and their addresses.

4.2 Collapsing Reported Off-site Facilities

There are approximately 3 million off-site facility records in TRI. However, many of these facility records actually represent the same facility; they are just reported in slightly different ways by the facilities transferring chemicals to them. In addition, approximately 1 million records are blank or not viable records. In order to make the geocoding process more efficient, it is necessary to first collapse the list of all reported off-site facilities into possible unique facilities. The first collapsing procedure removes all records that are not viable along with all of the records that are exact duplicates. This first stage collapses the off-site facility records from approximately 3 million to approximately 300,000.

Further collapsing, using algorithms in SAS to match addresses where the content is the same but the form is different (i.e., St. instead of Street), can bring the count down to approximately fifty thousand. However, the risk with this second collapse is in matching records that aren't exactly the same, and also in picking one address form to represent that facility, where another form might be better for geocoding purposes. Therefore, to decrease potential error in geocoding unique facilities, the almost 300,000 facility address records were sent to the geocoding service.

4.3 Geocoding the Off-site Facilities

TCS evaluated the 300,000 off-site facility address records. Their geocoding efforts resulted in a 50% street address match; 0.18% ZIP+4 centroid match; 0.16% ZIP+2 centroid match; nearly 47% ZIP code centroid match; and nearly 3% unmatched records. At this point in the process, this numbers may be misleading, since many of the 300,000 facilities are duplicates. Presumably, some portion of the ZIP code matches and unmatched facilities have problematic street addresses that may be "corrected" by accepting the better data of some other record of the same facility.

4.3.1 Collapsing off-site facilities after geocoding

A "fuzzy" matching SAS program (FIND_UNIQUE.SAS) was used to identify additional duplicate records that belong to a single unique facility. The term "fuzzy" refers to logical systems that do not require exact equality of two values in order to classify the two values as equal. In the name matching application, FIND_UNIQUE.SAS assigns two records to the same unique facility even if some identifying fields do not match exactly. This approach accommodates misspelled words and inconsistencies in how a facility might report its identifying information over time. For example, "DuPont," "Du Pont" and "E.I. DuDont" might all refer to the same facility. FIND_UNIQUE.SAS identifies a possible match based on similarity rather than exact equality in the name field and then decides whether to match the various spellings by examining the address fields.

Fuzzy matching always introduces the possibility of error. Two records may be matched that do not in fact belong to the same unique facility. Therefore, some discretion was applied in varying the program parameters and performing manual checks to balance two competing outcomes: a greater number of good/high confidence matches versus a greater number of erroneous matches.

The major parts of FIND_UNIQUE.SAS are:

1. Cleaning and conditioning the data;
2. Identifying a set of best names and addresses;
3. Matching records within the set of best names and addresses;
4. Finding indirect matches, where two records are matched not to each other but to a common third record.

The following sections describe in detail the SAS program and its application.

4.3.2 Cleaning and conditioning the data

The first part of FIND_UNIQUE.SAS corrects common spelling errors and inconsistencies and prepares the data for the matching algorithms. Data cleaning begins with the removal of extraneous characters, regularization of spaces and conversion of all letters to upper case. Then, words that occur frequently but do not aid in matching are deleted. These words include “COMPANY,” “LIMITED,” “POST OFFICE BOX,” “NOT AVAILABLE,” and numerous other words and their associated abbreviations and variations. If such words remain in the match fields, then a name such as “COB CORPORATION, P.O. BOX 2” would appear very similar to “AC CORPORATION, P.O. BOX 10.” The conditioning process converts the two names and addresses to “COB, 2" and “AC, 10," respectively.

Where relevant words commonly appear in various forms, the conditioning process substitutes a single form. For example, “NAT'L” and “NATIONAL” are both converted to “NATL.” A frequency analysis and visual review of words in the database led to some regularizations of facility names, such as “ADM,” “A.D.M.” and “ARCHER DANIELS MIDLAND,” or “EMPAC,” “EMPACK” and “EMPAK.”

For computational purposes, FIND_UNIQUE.SAS adds a leading blank and a trailing blank to each name and street address.

One example of how the conditioning might change a name field follows. If the reported name of a company is (the misspelling of “environmental” is intentional):

E	N	V	I	R	O	M	E	N	T	A	L		B	A	N	T	E	R			C	O	R	P	.

then the cleaned and conditioned version of the name would be:

	E	N	V	I	R		B	A	N	T	E	R	

The conditioning process concludes by correcting the state field when possible, based on the ZIP code field. FIND_UNIQUE.SAS does not assume that the ZIP code is correct whenever the reported state and ZIP code conflict. However, it does identify certain values in the state field as particularly susceptible to error. These suspect values were identified by checking reported state codes against reported city names and ZIP codes, using the *1996 World Almanac*. The conditioning process uses the state that corresponds to the reported ZIP code when the reported state is particularly susceptible to error or is not a valid state abbreviation. Table D-8 lists states that TRI reporters frequently misreport.

Table D-8
Suspect State Abbreviations

Reported State	Possible Actual State	Reported State	Possible Actual State
AR	AZ	MA	ME
AK	AR	MI	MS
AS	AR	MI	MO
CA	GA	MI	MN
IA	IN	MS	MO
IA	ID	NE	NV
II	IL	ON	OH
KT	KY	OP	OH
KU	KY	RH	RI
LA	AL		

Table D-9 lists state codes that were discarded in favor of the state corresponding to the reported ZIP code if and only if:

1. The reported state is listed in the "Reported State" column, and

2. The state corresponding to the reported ZIP code is the state listed in the "Possible Actual State" column of the same row.

For example, if the reported ZIP code is "85607" and the reported state is "AR," then the program corrects the state to "AZ." However, if the reported ZIP code does not begin with "85," then this section of the program makes no change to the state code.

Another section of the conditioning process corrects state codes in certain city name and state code combinations. For example, where the reported city name is "BALTIMORE" and the reported state is "MA," the SAS program changes the state to "MD." The program also changes Canadian province codes to "CN."

4.3.3 Identifying a set of best names and addresses

The purpose of the second part of FIND_UNIQUE.SAS is to reduce the number of records to be matched as quickly as possible. Since the time required to match all records in a dataset to each other increases exponentially as the number of records increases, it is important to perform preliminary matching using a simpler method where possible. FIND_UNIQUE.SAS does this by sorting records by facility name and comparing adjacent records. Thus, this early round of matching compares each record only to the preceding record and finds a match only in cases where the similarity is quite strong.

Specifically, the program sorts the data by the first ten non-blank characters in the facility name. If a reported facility name begins with the same ten characters as in the preceding record, the program compares the street addresses and ZIP codes and assigns three scores that measure the closeness of match in these location fields. If the scores exceed specified thresholds, then the program matches the two records to a single facility. Similarly the program then sorts the data by the first ten non-blank characters in the facility street address and compares the names in adjacent records.

FIND_UNIQUE.SAS calculates three scores that measure how well two names or two street addresses match. The definitions below use two new terms: *source* and *target*. The *source* is the set of words – i.e., name or street address – for which a match is sought. The *target* is the set of words that is being compared to the source. In the current part of the program, which compares adjacent records only, it does not matter which comparison value is designated the target and which is designated the source.

1. **<u>Match</u> <u>Score</u>**: The match score is the weighted proportion of letter pairs in the source also found in the target. A score of 0 means that no letter pairs in the source occur anywhere in the target. A score of 1 means that 100 percent of the letter pairs in the source also occur at least once in the target.

   Example:

Source =		B	A	N	T	E	R		.		
Target =		B	A	N	D	A	I	D	S		.

   The eight letter pairs in the source are: _B, BA, AN, NT, TE, ER and R_, where "_" represents a blank. Of these, _B, BA and AN also appear in the target. Therefore, the unweighted match score is 3/8 or 37.5 percent.

   FIND_UNIQUE.SAS introduces variable weights to allow the user to apply expectations about where errors are most important. In the current application, weights for letter pair matches decline exponentially so that matches near the beginning of the target are more valuable than later matches. The use of this model was based on an informal examination of the data.

2. **Position Score**: The position score measures similarity in the sequencing of letter pairs. The reason this is important is that the match score gives credit for a letter pair match regardless of where the letter pair occurs. In the above example, if the target had been “AFTER BANDAIDS,” the match score would have increased to 7/8 or 87.5 percent because the letter pairs TE, ER and R_ occur in “AFTER.”

The position score depends on where a matched pair is with respect to the first matched pair. In the following example, the first pair matched is _B, which occurs in the target at position 7.

Source =		B	A	N	T	E	R	
Position =	1	2	3	4	5	6	7	8

Target =		A	F	T	E	R		B	A	N	D	A	I	D	S	
Position =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

The position score is similar to the root mean square (RMS) algorithm commonly used to measure error in diverse situations. A position score of 0 indicates that the matched letter pairs occur exactly in order and at the same relative positions in both the target and source. Higher scores indicate poorer matches.

3. **Leftover Score**: The leftover score measures the percent of the target that is not matched to any letter pairs in the source. The leftover score helps compensate for the tendency of the previous two scores to overmatch short sources. To illustrate, in the following example, the match score is 100 percent and the position score is 0 – both optimum values.

Source =		B	A	N		.				
Target =		B	A	N	D	A	I	D	S		.

The leftover score measures the percentage of the target that is not matched to any letter pairs in the source. As in the match score, the leftover score uses a weighting system to give more weight to letter pairs that are most useful in discriminating between spelling variations and non-matching names. The best value is a leftover score of 0, and the worst value is 100 percent.

The comparison of adjacent records ends with one more iteration: by five digit ZIP code. The first two iterations examine records sorted by ten characters of the name and then by ten characters of the street address. The ZIP code iteration sorts all the records by five digit ZIP code and then compares adjacent records within each ZIP code for goodness of fit in both the name and street address fields.

FIND_UNIQUE.SAS allows the user to specify separate threshold values for each score and for each match field. In the iteration where names begin with the same ten non-blank characters, the thresholds for street address matches are relaxed slightly when five digit ZIP codes match exactly.

4.3.4 Matching records within the set of best names and addresses

The most powerful part of FIND_UNIQUE.SAS compares each record within a dataset to every other record, but it is also the slowest. For this reason, it is important to use Part II first to match closely-related records through comparisons of adjacent records.

Part III simultaneously scores and evaluates four match fields: name, street address, state and ZIP code. The program compares each record (source records) to all other records (target records). If the source record matches multiple target records, then the source is assigned to the target with the most frequently reported identifying data.

For example, assume that all of the following records match each other and they are all in the same state:

Name	Street	ZIP	Frequency
BANTER	10 MAIN ST.	12345	10
BANTER	P.O. BOX 40	12345	2
BATNER	10 MIAN ST.	72345	1

The “Frequency” column indicates how many times each version of the identifying data occurs in the database. Ten times, the facility reported its name as “BANTER,” its street address as “10 MAIN ST.” and its ZIP code as “12345.” Since this combination of identifying information occurs more frequently than the other two, FIND_UNIQUE.SAS assigns “BANTER,” “10 MAIN ST.” and “12345” to all thirteen records.

As part of this step of the program, the data are exported to an Excel spreadsheet, where some manual matches and corrections supplement the SAS matching. The data are then imported into SAS again, where processing continues.

4.3.5 Finding indirect matches

In the final part of FIND_UNIQUE.SAS, the program consolidates all the information about matching records and finds a set of unique facilities. In particular, Part IV finds indirect matches, where record A matches record B and record B matches record C but a comparison of A to C fails the goodness of fit thresholds. In this case, A and C should be matched even though they fail in the direct comparison.

In the following hypothetical example, the first record might match the third record by a four letter-pair match in the name field (_B, BA, ER and R_) with an optimal position score of 0, combined with an exact match in the street address field and an exact five-digit ZIP code match.

Name	Street	ZIP
BANTER	10 MAIN ST.	123450040
TRENTON PLANT	P.O. BOX 40	12345
BATNER TRENTON PLT	10 MAIN ST.	12345

The second record might also match the third record based on good match and position scores in the name field and an exact match in the ZIP code field. Therefore, all three records pertain to a single unique facility, even if the first and second records might fail to match using a direct comparison.

4.3.6 Identifying and assigning the best state match

The fuzzy matching program results in two output files: (1) the original file of offsite facilities in which each observation is labeled with the identification number ("ID_MATCH") of a unique off-site address for which it matches (approximately 3 millions records), and (2) a file which represents the legend of unique off-site records based on the ID_MATCH identification number (39,279 for Reporting Year 2000). The latter file contains the records used in the display of off-site facility information in the RSEI model, such as the best name and address or locational coordinates determined from earlier routines of the fuzzy matching program. However, this unique addresses file does not output the best state associated with each facility as it does for name, street, city, and zip.

To develop a state value for each of the unique off-site addresses, the 39,279 facilities were plotted to retrieve the state in which they mapped to. Similarly, the state corresponding to the best zip value was also retrieved (i.e. BEST_ZIP as determined by the fuzzy matching program). A separate analytical routine was then performed in SAS to determine the BEST_STATE value. This analysis required the following preparatory procedures:

1. The original file of approximately 3 million reported off-site facility records was sorted based on the unique off-site identification number it was assigned;
2. The frequency of the reported state within each ID_MATCH group of records was calculated;
3. The state most frequently reported for each ID_MATCH group was retained.

As a result of these procedures, three different fields containing various state values could be compared for each unique off-site facility: the plotted state, the state corresponding to the BEST_ZIP, and the state reported with the highest frequency. The following rules were applied in their comparison and in the determination of the final BEST_STATE value:

1. If the plotted state = BEST_ZIP state = reported state, then the state was considered valid;
2) Alternatively, if any two of the three fields matched, then that state value was used;

3) Finally for instances in which the latitude or longitude = 0 or was blank, no plotted state could be determined, so the reported state, if available, was used.

Of the 39,279 unique addresses that resulted from running Hsing Min's collapse program, all but 119 resolved with a BEST_STATE based on this methodology. The remaining 119 off-site facilities where exported into Excel and manually evaluated since the three state fields were in disagreement. The three state fields were used as a guide and provided context during this manual verification of BEST_STATE. Some of the reasons for how the BEST_STATE was determined for these records, included:

1) Some combination of city/zip/state was confirmed on www.usps.com;
2) The state was in the facility name;
3) Searched on some combination of the name/street address/city on www.google.com for an exact match.

Among these 119 records were some facilities for which the lat/long coordinates were deleted. Reasons for deleting lat/long coordinates included: (1) they were erroneous (e.g., the facility was actually located in the UK or Canada, or the search on name and address revealed a different state that was NOT adjacent – if the state was adjacent, the lat/long was not deleted), or (2) no supporting data to make a determination could be found using any of the above mentioned methods. Finally, only three records resulted with no state value at all; and those lat/longs were also deleted because two were in the UK and one was located in Canada. These 119 records were then re-appended to the larger off-site address file resulting in the complete set of 39,279 unique off-site addresses.

4.3.7 Results

The geocoding procedure and the SAS algorithms collapsed the number of off-site records from the initial 3 million to a final set of 39,279 records. As shown in Table D-9, approximately 36 percent of the unique facilities were matched to high-quality street addresses. Note that each unique address may represent multiple reports of off-site transfers from multiple Form R's.

Table D-9
TCS Off-site Geocoding Match Results after Collapse of Duplicate Records

Coordinate Matches	Number of Records	Match Percentage
Street address (including hand matches)	14,000	35.64%
ZIP+4 centroids	93	0.24%
ZIP+2 centroids	71	0.18%
5-digit ZIP code centroids	23,614	60.12%
Unable to geocode	1501	3.82%
Total Unique Facilities	**39,279**	